AD-774 985

CONTROL CONCEPTS IN A SPEECH UNDERSTANDING SYSTEM

Paul D. Rovner, et al

Bolt Beranek and Newman, Incorporated

Prepared for:

Advanced Research Projects Agency

December 1973

Unclassified
Security Classification

AD 774 985

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)

Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, Massachusetts 02138

2a REPORT SECURITY CLASSIFICATION

Unclassified

2b GROUP

3 REPORT TITLE

CONTROL CONCEPTS IN A SPEECH UNDERSTANDING SYSTEM

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5 AUTHOR(S) (First name, middle initial, last name)

Paul D. Rovner
Bonnie L. Nash-Webber
William A. Woods

6 REPORT DATE	7a TOTAL NO OF PAGES	7b NO OF REFS
December 1973	~~18~~ 26	5

8a CONTRACT OR GRANT NO	9a ORIGINATOR'S REPORT NUMBER(S)
DAHC15-71-C-0088	BBN Report No. 2703 AI Report No. 9
8b PROJECT NO	
c	9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d	none

10 DISTRIBUTION STATEMENT

Distribution of this document is unlimited.

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY
	Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, Virginia 22209

13 ABSTRACT

Automatic speech understanding must accomodate the fact that an entirely accurate and precise acoustic transcription of speech is unattainable. By applying knowledge about the phonology, syntax, and semantics of a language and the constraints imposed by a task domain, much of the ambiguity in an attainable transcription can be resolved. This paper deals with how to control the application of such knowledge. A control framework is presented in which hypotheses about the meaning of an utterance are automatically formed and evaluated to arrive at an acceptable interpretation of the utterance. This design is currently undergoing computer implementation as a part of the BBN Speech Understanding System (SPEECHLIS).

i.

26

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Security Classification

A-31409

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Automatic speech understanding						
artificial intelligence						
computational linguistics						
SPEECHLIS						
speech recognition						
speech						
intelligent system organization						
speech understanding						

BBN Report No. 2703
AI Report No. 9

CONTROL CONCEPTS IN A SPEECH UNDERSTANDING SYSTEM

PAUL ROVNER
BONNIE NASH-WEBBER
WILLIAM WOODS

DECEMBER 1973

This research was supported by the Advanced Research Projects Agency of the Department of Defense under ARPA Contract No. DAHC15-71-C-0088 and ARPA Order No. 1697.

ib.

TABLE OF CONTENTS

1. INTRODUCTION

Speech understanding, whether done by people or by computers, demands great funds of knowledge. This is due to the inherent imprecision, variability and complexity of the acoustic signal into which all speech is encoded [1]. For example, the encoding of a word, spoken in running conversation, will be affected by its environment (the words surrounding it), its importance to the message (stress and intonation), its speaking rate, and its speaker. It is an apparent circular dilemma of acoustic-phonetics that one cannot precisely identify contextual effects without first identifying the context. However, we believe that by applying other sources of knowledge to an initially uncertain and imprecise acoustic transcription, in order to hypothesize possible higher contexts, this circularity can be broken.

In this paper, we are concerned with the problem of how to control the application of various sources of knowledge to this problem. A framework of concepts, data objects, queues, and programs is presented in which strategies for forming and evaluating hypotheses about the meaning of an utterance may be implemented and studied. One such strategy is described, and an example of its performance is given. A Speech Understanding System (SPEECHLIS) being developed at

Bolt Beranek and Newman provides the environment for this work, and derives much of its structure from this control framework. Though it is not our purpose here to discuss in detail the design of SPEECHLIS, it will be useful to the reader to know that it contains several knowledge sources as components -- acoustic-phonetic, phonological, lexical, semantic, syntactic, and pragmatic.

A listener does not just passively accept speech: he actively uses all his knowledge to structure uncertain and incomplete cues from the acoustic signal into a grammatical, meaningful and appropriate utterance. Sources of knowledge which are available to a listener include:

1. The acoustic-phonetic properties of the language -- Knowledge of the correspondence between physically varying parameters of the speech signal and the basic phonetic elements of the language (phonemes).
2. The vocabulary -- One presumes that an English utterance will consist of a sequence of English words, interspersed, perhaps, with pauses and non-speech sounds. A vocabulary constrains the possible sequences of speech sounds and the set of words which might fit a particular sequence.
3. Phonological rules -- These rules specify allowable or characteristic variations in the pronunciation of words or phonemes in particular environments.

4. The syntactic structure of the language -- A sound sequence which is heard as the word sequence "in other samples" will not be heard as "in of a samples", since the latter is ungrammatical.
5. The set of concepts and relations that are meaningful to the listener -- A sound sequence which is understood as "close the doors" will not be understood as "close the daws", since birds don't close.
6. Pragmatic considerations (knowledge of the current context or situation) -- A similar sound sequence may be heard as "close the Dewers" in a room in which the only thing open is a bottle of scotch.

Much of the above knowledge is specific to a problem domain. For our automatic speech understanding effort at BBN, we have chosen the task area of an existing natural language question-answering system (LUNAR) for the Apollo 11 moon rocks [2], which answers questions such as:

> How many breccias contain more than 10% anorthosite?
>
> In which samples was titanium found?
>
> Give me all references to olivine twinning.

In doing so, we have been able to draw upon our knowledge of lunar geology and question-answering system characteristics developed during work on that system. The LUNAR system operates with a lexicon of around 3500 words. As of this writing, SPEECHLIS is operating with a 250 word lexicon, with a larger one of about 1500 words in preparation.

2. OVERVIEW OF THE CONTROL FRAMEWORK

2.1 Data Objects

The control framework that we will discuss assumes the existence of programs which have access to various sources of knowledge. For example, acoustic-phonetic and phonological programs operate on a digitized wave form to produce an acoustic transcription of the utterance in the form of a collection of SEGMENT descriptors. By a segment we mean a portion of the utterance which is hypothesized to be a single phoneme. Each segment has a description which could in principle specify the phonetic identity of the segment, but in general merely constrains this identity to one of several phonemes. Alternative hypothesized segments may overlap in the utterance, resulting in a lattice of segment descriptors rather than a single string. Figure 1 gives an example of such a SEGMENT LATTICE. This structure allows for the representation of uncertainty or ambiguity both in the identity of a segment and in the determination of the segment boundaries.

Lexical retrieval and word matching programs are available to map sequences of segment descriptions into words. They do this by matching PHONETIC SPELLINGS of the words in the vocabulary against sequences of adjacent segments. The correspondence between a single phonetic spelling of a word and a segment sequence is called a WORD

MATCH. Since the acoustic transcription may make errors in the detection of segments, word matches involving missing or extra segments may also be made. The quality of the match is one indication of the likelihood that the word actually appears at that place in the utterance. Word matches to be examined by syntax, semantics and pragmatics programs are entered into a WORD LATTICE. (Such a lattice is illustrated in Figure 2.) In this figure, for example, the word "mean", spelled phonetically [min], or to use our computer representation [M IY N], matches from 2 to 5 in the lattice, while the word "any", spelled [ɛni] or [EH N IY], matches from 3 to 6.

Each phoneme in the above two spellings satisfies exactly the phoneme description of its corresponding segment. We do not assume however that the correct phonemic identity of a segment will always be among the set of phonemes postulated by the acoustic-phonetic and phonological programs. Rather we assume that if they err, the correct phoneme will be similar in acoustic characteristics to those given. For example, at the beginning of the segment lattice, the first two phonemes of the word "give", spelled [gIv] or [G IH V], match the segment descriptors perfectly. The third, [v], is sufficiently close to [b] acoustically, that a word match is made for "give" and entered into the word lattice. However, since the acoustic transcription is the best evidence we

have of what the utterance was , our confidence in "give" actually beginning the utterance is less than if each of its phonemes had matched perfectly.

Interacting with the word lattice, the higher level components of the system (syntax, semantics and pragmatics) form internal data objects called THEORIES representing hypotheses about the original utterance. A theory contains a non-overlapping collection of word matches which are postulated to be in the utterance, together with syntactic, semantic and pragmatic information about this collection and scores representing the evaluations of that theory by various knowledge sources.

Theories grow and change as additional bits of evidence for or against them are found. A principal mechanism for accomplishing this is the creation of MONITORS. A monitor is a trap set by a hypothesis on new information which, if found, would result in a change or extension of the monitoring hypothesis. However, the reprocessing that is called for when a monitor is noticed is not done immediately. Rather an EVENT is created, pointing to the monitor and the new evidence. This event is evaluated to decide if and when to do it.

In addition to waiting for new information (by setting monitors), the higher level components can actively seek it out. One way this is done is by PROPOSALS. A proposal is a request to match a particular word or set of words at some point in the utterance: any of the higher level components can make proposals.

A short example should illustrate the above concepts more clearly. Notice the robust word match for "chemical" in the word lattice shown in Figure 2. The semantics component knows about CHEMICAL ANALYSES and CHEMICAL ELEMENTS, but not about CHEMICAL as an independent concept. Since "chemical" matches well, semantics might postulate that one of these concepts is being designated. It could propose "analysis", "analyses", "determination"(all naming the first concept) and "element", requesting them to be compared against the segment lattice, right adjacent to "chemical". Since "analyses" and "analysis" match well, events would be created, linking the hypothesis for "chemical" with those for "analysis" and "analyses". Given that CHEMICAL ANALYSIS refers to the amount of each major element in some rock, e.g. "chemical analyses of fine-grained lunar rocks", any hypothesis created for "chemical analyses" will monitor for an instantiation of the concept ROCK. If found, it will give additional support to the theory that what is being discussed is indeed the chemical analyses of some rock.

2.2 Evaluation Mechanisms

A notion central to the control framework is that of evaluation: one cannot afford to spend time on activities unlikely to produce good results. The various scores associated with a theory are used by Control to allocate its resources to where it expects to achieve results. In this section, we discuss how knowledge is brought to bear in computing these scores.

The score of a word match depends on how well each of the phonemes in the phonetic spelling matches the corresponding sound description in the segment lattice. Among the factors taken into account in making this match are such things as:

1. A priori information about the similarity of sounds (e.g. [i] is more similar to [I] than to [a].)

2. Cues from comparing the actual duration of a segment with duration information derivable from the phonetic spelling using vowel tenseness and stress.

3. The likelihood of missing or extra segments. This is determined both from empirical studies of the segmentation errors which are made by the acoustic-phonetic programs and from phonological rules which indicate the sounds in each phonetic spelling which are likely to be missing or extra.

4. The length of the word. Long words which match well get a boost in score because it is relatively unlikely that good, long word matches would be detected at random.

The score of a theory is a weighted sum of its lexical, syntactic, semantic and pragmatic scores. The lexical score depends on the average word match score for the words in that theory, the number of adjacent word matches, and acoustic effects at their boundaries. The semantic score is based on an evaluation of the conceptual structures that semantics has built, reflecting whether they are complete or lack some obligatory component. In the latter case, semantic confidence in the theory is lowered.

The syntactic evaluation is based on the ability to assign syntactic structure to the hypothesis. Using an augmented transition network grammar [3] and a parser capable of working with disjoint sequences of word matches, the syntactic component tries to parse each such sequence and decide whether sequences could be joined into a larger syntactic structure. If a word match sequence fails to parse, or if two nearby sequences cannot be bridged in any way, syntactic confidence in the hypothesis will be low.

Currently, SPEECHLIS contains very limited pragmatic knowledge: only the most rudimentary speaker and context models are available for use in evaluating a theory. Observing the relationships postulated by syntax and semantics, the pragmatic component evaluates the likelihood of an utterance that would contain them. For example, in the context of question-answering, questions and commands are more likely than statements: so pragmatics looks for syntactic evidence of sentence type in making its evaluation. The question-answering context also makes certain semantic concepts more likely than others. For example, the concept of the machine giving the user something or of the user needing something is more likely to be expressed than any particular concept, such as that of spectrographic analysis. The pragmatic component uses the conceptual structures that semantics has built to evaluate their likelihood of occurrence. (This evaluation is currently user independent, but we expect eventually to deal with a dynamically developed model of the user's interest.)

There is a further evaluation based on the consistency of the semantic and syntactic structures. Associated with each conceptual structure that semantics has built is a condensed description of the ways in which that structure might be realized syntactically. If none of the structures that syntax can build correspond to these , this discrepency lowers the likelihood of the theory actually representing

part or all of the original utterance.

An event is evaluated in the same way as a theory: that is, the score of an event will reflect the score of the suggested new theory.

2.3 A Control Strategy

Within the framework of word matches, theories, evaluation mechanisms, etc., a preliminary control strategy is undergoing computer implementation. In this strategy, the proposals, theories and events that occur during processing are evaluated and placed on three separate queues, ordered by the scores of their elements. The basic characteristic of this strategy is to select elements from the tops of these queues and process them.

The first activity of the control programs is to call the acoustic-phonetic and phonological programs to construct an initial segment lattice from the speech signal. A word lattice of robust word-matches is then constructed by a program which scans the segment lattice with the aid of the dictionary. In addition, a set of words which are pragmatically likely to begin an utterance are matched at the beginning of the segment lattice. As each such word match is found, it is entered into the word lattice and given to the semantic component for analysis. If the word has semantic content, a theory is created for the word

match, designating all semantic contexts in which it could appear. If a monitor is noticed indicating that a word fits into the semantic context of a theory which was created earlier, an event is created which associates the new word match with the old theory. Proposals for specific content words which are likely to appear adjacent to the new word match are created and added to the proposals queue.

For each new word match, appropriate inflexional endings and auxiliary verbs are matched against the segment lattice and associated with the word match if they match well.

After the initial set of robust word matches are examined, the proposals that are likely to be productive are processed, thus introducing new word matches and triggering a new round of semantic analysis. The events at the top of the event queue are then handed back to the semantic component for further processing. For each event, a new theory is created with a modified semantic context and entered into the theory queue. This may result in additional events, as semantics notices other word matches in the word lattice which fit into the modified context. In this way, semantics assembles meaningful sets of content words.

As new theories are created, each is examined to determine whether it might be fruitful to call upon

syntactic knowledge to develop further support for it. Since the number of possible parsings decreases with the number of adjacent or "close" word matches, this decision is made on the basis of the number of adjacent word matches in the theory, the size of the gaps between word match sequences, and the absence of content words in the word lattice which would be added to the theory by semantics.

Syntactic knowledge is used to postulate grammatical structures that may obtain among the words in a theory. For example, for "... people done chemical analyses...", syntax could suggest that "people" is the subject of the verb "done", "chemical analyses" is the noun-phrase object, and that an auxiliary verb appears somewhere in the utterance (probably at the beginning) to modify the past participle "done". Such grammatical information is checked for consistency with the postulated semantic structures, to determine for example whether it makes semantic sense for "people" to do something. Function words (e.g. determiners and prepositions) which are likely to appear adjacent to a sequence of word matches are proposed by syntax in the context of these grammatical structures, and added to the theory as a refinement if they are found. Each small gap between sequences of word matches is analyzed, and a strong attempt is made to find a small word which fits. If none is found, it is likely that one of the word matches adjacent to the gap is wrong.

3. AN EXAMPLE

To illustrate the operation of the above control strategy, we will consider a specific example. The segment lattice shown in Figure 1 was constructed by hand from a speech spectrogram during a study of human performance in spectrogram reading experiments [4]. The word lattice shown schematically in Figure 2 was constructed from it by looking for robust word matches and possible adjuncts (inflections and auxiliaries) and by trying to match pragmatically likely words in sentence initial position.

Following this first pass in which word matches were entered in the word lattice and given to semantics for processing, there were 42 theories and 48 events. (Some pruning was done to eliminate unlikely events.) The five events at the top of the event queue were ones linking "chemical" and "analyses", "modal" and "analyses", "chemical" and "analysis", "modal" and "analysis", and "metal" and "analyses". (One can analyze a rock for its metal content.)

Processing these five events led to the creation of five new theories and 55 new events. At this point, the best events called for linking:

1. "give" (initial position) and "chemical analyses"
2. "give" (initial position) and "modal analyses"
3. "give" (initial position) and "chemical analysis"
4. "print" (initial position) and "chemical analyses"
5. "have" (initial position) "done" and "chemical analyses"

Notice that the top four events were quite reasonable though incorrect. Five new theories and 20 new events were created during this round of processing.

The next round of event processing brought the following five events to the top of the queue:

1. "have ... done chemical analyses" and "people"
2. "have ... done chemical analyses" and "rock"
3. "give ... chemical analyses" and "me" (following "give")
4. "give .. chemical analyses" and "us" (following "give")
5. "give ... chemical analyses" and "I" (following "give")

Notice that the top two events were each filling up a different semantic role in the concept of doing a chemical analysis - the agent of the doing and the object of the analysis. As to the "give I" event, semantics does not know

that this is syntactically incorrect. Again five new theories were created during this round, but these resulted in only the five events shown above.

At the start of the fourth round of event processing, the five best events were:

1. "have ... people done chemical analyses" and "rock"
2. "have ... done chemical analyses ... rock" and "people"
3. "give me ... chemical analyses" and "rock"
4. "give us ... chemical analyses" and "rock"
5. "give I ... chemical analyses" and "rock"

Notice that the top two events would result in the same theory. However, before a theory is created, the control strategy checks that no such theory already exists. If one does, processing is halted on that event so that duplication does not occur. (However, this ability to arrive at the same theory from several directions is necessary since it allows us to put together incomplete structures, irregardless of which pieces are missing.) The four resulting theories were semantically complete: both agent and object of "doing" had been identified, as had the object of "chemical analyses", and agent, recipient and object of "give". At this point, semantics could not contribute anything to these good theories, and they were sent off to syntax.

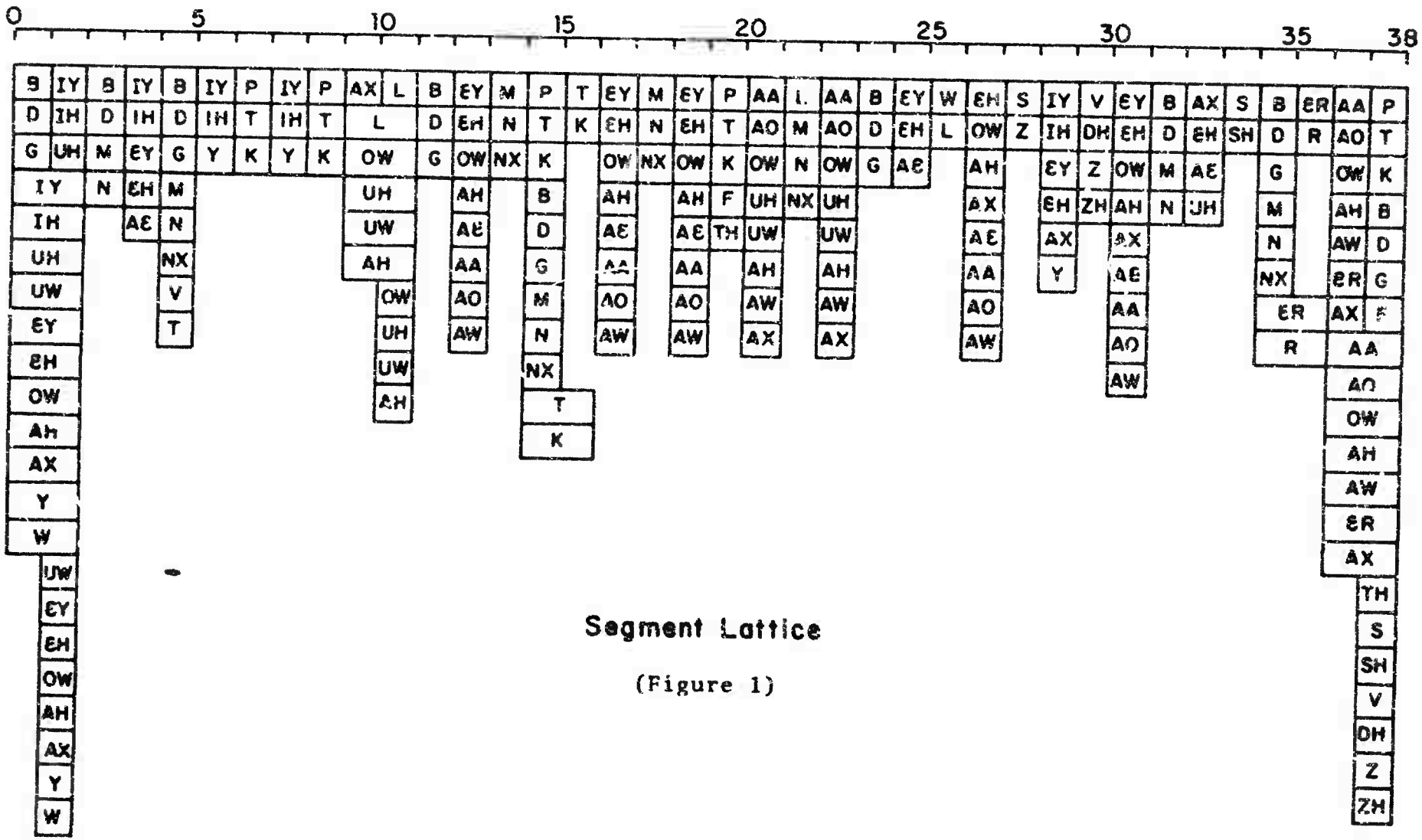
Syntax noticed the determiner "any" in the word lattice which could precede "people" syntactically, and it created an event which would refine the first theory with the word match for "any". In addition, syntax proposed determiners before "rock", since none occurred in the word lattice. This and additional proposals brought word matches for "this" and "in" into the word lattice. These were added to the theory by syntax, resulting in a semantically meaningful, grammatically correct one which spanned the utterance. This was, at the time, sufficient criteria for accepting the theory "Have any people done chemical analyses on this rock" as a correct understanding of the utterance.

4. CONCLUSION

Both the control framework and strategy presented above are incomplete since many problems have still to be faced. Our most difficult current problem involves recognizing the state when the system is just thrashing around, when no theory deriving from our current strategies is going to emerge as a good candidate for the whole utterance. We need to use our knowledge sources to decide which pieces of existing theories are most reliable, and which pieces should be tossed out. To get a better feeling for the possibilities, we expect to use the technique of "incremental simulation" [5], in which a person simulates a part of the system which is not yet formulated to gain insight into how it might work.

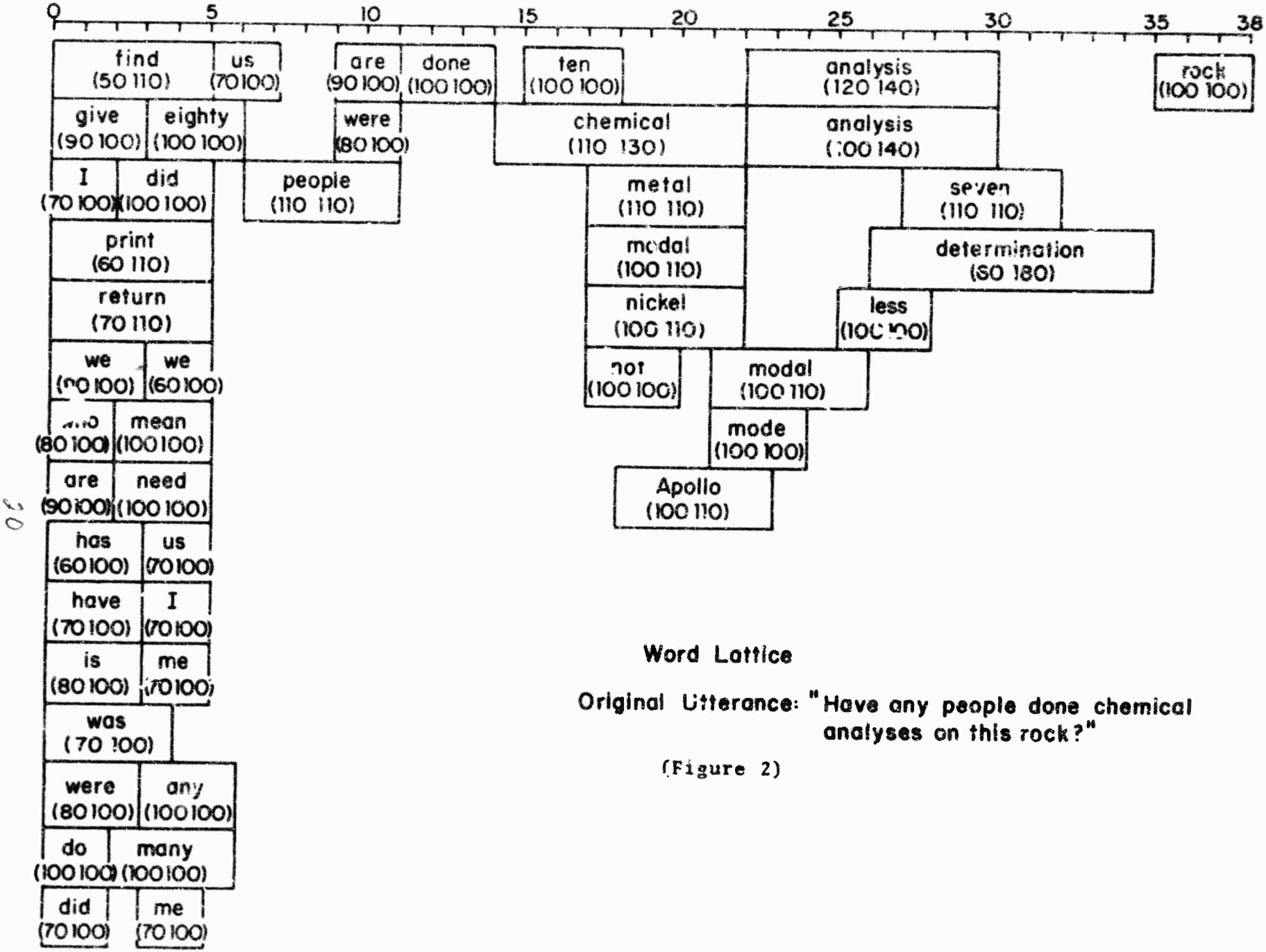
Another pressing problem is the need for a more rigorous foundation for measuring confidence in evidence and combining such measures into measures of confidence in theories and events. As complexity increases, our current methods will become more difficult to manage.

Other problems will arise from our imminent transition to a larger vocabulary and projected transition into different task domains. The attempt to solve all these problems will test the adequacy of our control framework in dealing with a world of uncertain and incomplete information.



Segment Lattice

(Figure 1)



Word Lattice

Original Utterance: "Have any people done chemical analyses on this rock?"

(Figure 2)

REFERENCES

[1] P. Denes and E. Pinson, The Speech Chain, Bell Telephone Laboratories, Murray Hill, New Jersey, 1963.

[2] W.A. Woods, R.M. Kaplan, and B. Nash-Webber, The Lunar Sciences Natural Language Information System: Final Report, BBN Report 2378, Bolt Beranek and Newman, Cambridge, Mass., June 1972.

[3] W.A. Woods, Transition Network Grammars for Natural Language Analysis, Communications of the ACM, Vol. 13, No. 10, October 1970, pp. 591-602.

[4] D.H. Klatt and K.N. Stevens, Sentence Recognition from Visual Examination of Spectrograms and Machine-Aided Lexical Searching, Conference Record, 1972 Conference on Speech Communication and Processing, Newton, Mass., April 1972.

[5] W.A. Woods and J. Makhoul, Mechanical Inference Problems in Continuous Speech Understanding, Proceedings of the Third International Joint Conference on Artificial Intelligence, August 1973, pp. 200-207.